National College of Ireland

# Role of Social Media in Start-ups Success using Machine Learning Approaches: Twitter

MSc Research Project
Data Analytics


Anusha Ananth
Student ID: x21131465

School of Computing
National College of Ireland


Supervisor: Mr Vladimir Milosavljevic

**National College of Ireland**
**Project Submission Sheet**
**School of Computing** https://www.overleaf.com/project/639936cae

| | |
|---|---|
| **Student Name:** | Anusha Ananth |
| **Student ID:** | x21131465 |
| **Programme:** | Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Mr Vladimir Milosavljevic |
| **Submission Due Date:** | 01/02/2023 |
| **Project Title:** | Role of Social Media in Start-ups Success using Machine Learning Approaches: Twitter |
| **Word Count:** | 5917 |
| **Page Count:** | 19 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 1st February 2023 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Role of Social Media in Start-ups Success using Machine Learning Approaches: Twitter

Anusha Ananth
x21131465

**Abstract**

The rise of start-ups has had a significant impact on the economy's growth. A start-up grows through creativity, partnership, publicity, and so on. Start-ups utilize social media platforms to publicize their ideas and goods as a result of the internet explosion over the last decade. A business is referred to as a "unicorn" if it surpasses the billion-dollar milestone. Twitter, WhatsApp, Facebook, Instagram, and other popular social media networks. Sometimes financing is reliant on client tweets or feedback. The major purpose of this research is to ascertain the influence of tweets on start-up investment. The project's principal purpose is to develop machine learning and data mining algorithms that can be used for large start-up funding datasets to gain insights that will assist venture capitalist specialists in offering timely insights regarding series investments. A hybrid strategy of feature selection utilizing the K-Fold Statistical technique, regression, and neural network approaches was built to identify relevant factors for the forecast of start-up series funding. For each model, the Root Mean Squared Error is generated and used to evaluate its performance. The hybrid technique extracted 30 funding-related characteristics from Twitter data, and the Recursion Neural Network (RNN) produced a Root Mean Squared Error (RMSE) of 0.65 by outperforming the other models.

## 1 Introduction

Customer feedback and satisfaction have risen to prominence in the twenty-first century. This input is then shared on social media platforms including Twitter, YouTube, Facebook, Instagram, Snapchat, Pinterest, Netflix, Google, and others. Every minute, around 87,500 individuals tweet on Twitter. Every minute, 4.5 million individuals are viewing YouTube videos on the internet. Every minute, Google receives over 3.8 million search inquiries. So, every minute, a client sends feedback or expresses an opinion, and the amount of time individuals spend on the internet has risen over the last decade.

The economic system has evolved in technology in the twenty-first century, and every day new technology and new business are formed. These firms achieve success within the first five years of entering the market. And those who exceed a billion US dollars are referred to as "unicorns." In any developed or developing country, entrepreneurship has become the most essential indicator of a country's economy.

A start-up is a business in its early stages. The success of a start-up may be judged in a variety of ways. A start-up grows with money from venture capitalists or angel

investors. Alternatively, start-ups are purchased by large tech corporations, which may self-fund and expand a business to enormous heights, or attain an IPO and flourish as a company on their own. In recent years, start-ups have received a great deal of attention. Since the COVID-19 epidemic (Bangdiwala et al.; 2022a), knowledge of start-ups and entrepreneurship has risen dramatically and massively. In the United States, there were around 5 million applications to form a business in July 2020, as compared to 2019. The number of applications filed increased by 95%.
Start-ups have now become a solution to every country's economy. They are also giving a new source of revenue and assisting individuals in making a life throughout the epidemic. And, since the pandemic, the start-up market has been hard and competitive, adding dynamism to the economic structure.

However, not all start-ups thrive; in fact, there are more failures than successes. According to estimates, 90% of new businesses failed to succeed in 2019. And around 21.5% of start-ups fail to succeed in their first year in the market, 30% fail to succeed in their second year, 50% fail to succeed in their fifth year, and over 70% fail to succeed in their tenth year (Bangdiwala et al.; 2022a). As a result, it is difficult for today's entrepreneurs to get the guts to bring their start-ups back on track to success. And there is no such thing as a success manual or a certain road to success. It all relies on things such as the IPO, stakeholders, concepts, and how the product is marketed in the market.

## 1.1 Research Question

The research project's goal: - Is it possible to discover the effect of social media (Twitter) platforms on the success of start-ups through its funding rounds using machine learning algorithms?

## 1.2 Research Objective

**First Objective:** A correlation study to find the relation between the tweets and the funding rounds of start-ups.

**Second Objective:** Balancing of data and visualization of data, balancing of data before the processing of the data.

**Third Objective:** Implementation of Machine Learning models and evaluation.

**Fourth Objective:** Comparison of Models by RMSE value.

# 2 Related Work

The methodology employed in this case study endeavour is based on an algorithm that pinpoints the essential traits associated with a start-success. up's Particle Swarm Optimization has been used to choose features from among the many elements that affect a start-up's success. The features are picked to make it easier to assess the start-up's success. The framework offered an accuracy of 92.3% after being trained using a machine-learning classification algorithm. Support Vector Machine, Decision Tree, and Linear Re-

gression were the three methods employed. 52 out of 116 characteristics were also chosen for analysis (Pasayat and Bhowmick; 2021).

Predicting the success of crowdfunding is the focus of this work. Here, a novel architecture for the deep cross-attention network is put forward that makes use of data from project profiles' videos and written descriptions (Tang et al.; 2022). To improve upon the baseline projections, a thorough examination is conducted. To develop improved feature representation, cross-attention blocks are stacked on top of one another. Two datasets were employed, and the feed-forward approach was used for one of the datasets. They have covered the success of start-ups that are merging or being bought in this article and there are around only 10% of start-ups that succeed in 2019 (Bangdiwala et al.; 2022b). To determine whether the start-up will be integrated or not, five models are employed. They have employed Decision Trees, Random Forest, Logistics Regression, Gradient Boost, and MLP Neural Networks (DT). The models were trained using the important characteristics to determine the company's trajectory. The average accuracy for all the models was 92%.

This study established an association between business growth and bank loans (Åstebro and Bernhardt; 2003). Bank loans are a major reason why very tiny start-ups with higher sales fail. The source of money can occasionally have an impact on a start-up's success. If a bank approves the finance, the business's chances of survival are decent. However, if the start-up is funded by a family member, close friend, or other close relatives, its chances of survival are very slim. Instead, if the company owner obtains a loan while still having a mortgage on his home, the interest rate will be lower, and no evaluation will be performed before issuing the loan. Additionally, the business owner will be held accountable, which will motivate them to work hard to succeed. In this essay, the entrepreneur's education is highlighted, and the sort of load carried is taken into account to determine the start-up's success rate. Following the use of the regression model in this inquiry, several aspects were taken into account. It was shown that borrowing money from relatives and friends had a lower likelihood of resulting in company success.

There is a framework that will analyse the organizational influence on the tech start-ups as a result of the impact on tech start-ups (Gidron et al.; 2021). This framework was created for the formulation of sustainable objectives. The influence of tech start-ups is the focus of a machine learning model. Both investors and start-ups are major stakeholders in the development of these businesses. However, there is a lack of evidence to make this determination in a start-up's early phases. The likelihood of predicting with the use of machine learning algorithms is very high. Algorithms for machine learning perform better with larger datasets. CrunchBase was used to extract the datasets. The two algorithms employed are LightGBM and XGBoost (Yin et al.; 2021). And investors and start-ups benefit from these insights. The start-up's post-money is crucial since it is used to validate a variety of qualities that might be crucial to the success of start-ups. Bayesian optimization combined XGBoost with hyperparameter tuning (Ang et al.; 2022). A start-primary up's success may be judged by whether it is purchased or goes public. Additionally, 95% model accuracy was attained. This study offers start-ups, governments, and venture investors a new point of view.

One of the hardest things is predicting a company's performance. Approximately 213,171

businesses were considered for the prediction study (Żbikowski and Antosiuk; 2021). Additionally, there were three algorithms: Support Vector Machine, Gradient Boosting Classifiers, and Logistic Regression (SVM). And the logistic regression yielded results with a maximum accuracy of 87%. Additionally, K-fold statistical analysis was used to tune the hyperparameters. Additionally, start-ups from Asia, Europe, and the US were taken into consideration, and data from CrunchBase was gathered. There were 17 interviews (Hopkins and Booth; 2021) with stakeholders from various organizations done as part of this study. The data ethics were observed when analyzing the information from these interviews. To obtain reliable measurements, machine learning methods were also utilized. These have modest to medium-sized organizations that analyze machine learning. That will facilitate decision-making for organizations.

There is no suitable method to invest in the early phases of start-ups, thus angel investors are those who make investments in them. Venture investors and start-ups can communicate more effectively because of this issue. It will be possible to forecast the success of the start-up using these machine learning algorithms and the relevant information about the start-ups that are accessible on websites like Crunchbase and Kaggle. Additionally, this case study may be utilized as a tool for VCs to examine the first circumstances of a start-up (Arroyo et al.; 2019). In this instance, a multi-class machine learning classifier is used to train and analyse the datasets in order to produce better and more accurate findings so that venture capitalists may make better decisions when investing in start-ups in their early phases. For many venture investors, the investment structure of start-ups in their early phases continues to be a challenge. Additionally, a company's study for the 20-year time frame will be done for around 600,000 enterprises (Corea et al.; 2021). To provide angel investors with a clear notion of where to invest in start-ups in their early phases, about 21 factors were taken into account in the evaluation and machine learning algorithms. Seven software developers from various start-ups participated in this study (de Souza Nascimento et al.; 2019). This study examines how machine learning models are created by start-ups. It is stated that four separate phases are taken in this procedure to create a machine-learning model.

There aren't many start-ups that focus on machine learning and artificial intelligence. Furthermore, the healthcare industry has extremely few start-ups(Vijai and Wisetsri; 2021). Additionally, many underdeveloped nations do not have a robust health sector. The development of the country's economy will be aided by the numerous artificial intelligence start-ups working in the healthcare industry. Startups are now a viable economic answer in every country. They are also providing a new source of cash and supporting individuals in surviving the disease. Furthermore, since the pandemic, the start-up market has been difficult and competitive, providing vitality to the economic structure.

The success of start-ups is a topic of extensive research, and there are several traits and elements that influence start-up success. And the elements that aid in the analysis of the investment and the aspects that aid in the investment by venture capitalists and angel investors. Only seasoned investors can analyse these elements and characteristics that influence the investment. Investment angel education is never included in a conventional education program. Success for start-ups can also be determined by the milestones attained during the first five years of business in addition to the acquisition or initial public offering (IPO) (Veloso; 2020). The factors that determine the success of the start-up

will be analysed using machine learning algorithms in this work employing a step-by-step method starting from the start-up's early stages and continuing as it grows. Additionally, all currently employed machine learning algorithms will be evaluated for accuracy, and the top algorithm will be utilized to assist investors in making investment decisions.

Artificial intelligence (AI) is a tool that may be used in business to enhance operations and provide better judgments. Additionally, according to Crunchbase, there are about 27,900 startups as of September 2021 (Weber et al.; 2022). AI start-ups use cutting-edge techniques to address issues in the field. And with 84% accuracy, the issue has been resolved. Even if some issues are beyond the capabilities of artificial intelligence, humans are the only beings that can comprehend some concepts that are beyond the understanding of AI startups Start-ups focused on AI and machine learning emerged during the epidemic and were later considered in the UK service sector (Buchanan and Wright; 2021). Additionally, the influence of AI has spread to other sectors, and machine learning has had a big impact on areas like algorithm trading, credit scoring, and fraud detection. Utilizing data from the UK, these applications were analyzed. The research is then finished once the models' performance has been evaluated.

One of the key factors in a country's development is its industry and the new businesses that enter it and contribute to the growth of its economy. New firms, start-ups, and unicorns are highly significant in many wealthy and developing nations alike. And the nation's youth are mostly responsible for the growth of start-ups. There won't be economic progress in the nation if the kids aren't engaged in development. The university-age generation is where fresh ideas in youth originate. The fact that young people do not engage in the creation of start-ups must have some root cause. There is a dearth of entrepreneurial competency, particularly in emerging nations like India. The author of this study presents the hypothesis that Indian university students lack entrepreneurial ability. About 198 universities provided the information for this report (Sharma and Manchanda; 2020). Seven machine learning (ML) algorithms have also been developed to anticipate university students' lack of proficiency. The study had a favourable result, and the poll revealed that 67% of the male respondents supported the idea that a lack of entrepreneurial understanding implies a lack of awareness and knowledge. Additionally, because of increased academic pressure, it is harder to focus on such aspects of one's life.

It will be difficult to foresee a start-success up's manually, or for equity investors to make a worthwhile investment. Furthermore, it may be difficult to consider every factor that will influence a start-success. up's Furthermore, manual decision-making is time-consuming and inaccurate. When employing manual decision-making methods, data mining approaches, on the other hand, assist in taking into consideration all elements that are hidden from view. As a result, constructing a machine-learning model may free up time for start-up founders, venture capitalists, and angel investors. The major purpose of this technique is to determine the significance of Twitter tweets in the success of start-ups.

The case study of Theranos is discussed in the paper, where a lie can be left undisclosed in the early stages of the funding as it does not breach the law between the investor and the entrepreneurs. In the case of Theranos (Mayer; 2022), a project taken up was failing and the founders of the enterprise did not disclose the details to the public to continue to gain funding from the investors. At a point in time, the pot filled with lies overflowed and

could no more be kept as a secret. One of the major traits that were that an innovative project has a great affinity to failure and to maintain the project someone who is capable will have to handle it. And it is very hard for angel investors or venture capitalists to the happenings inside a company only a person from the enterprise can hide it from the investors. For any start-up to start a business with great novelty is difficult and it is difficult for the investors also to invest. And there is always a debate that whether start-ups should follow profitability or novelty (Lee; 2022), and whether there should be less hierarchy in the company. When there is lesser hierarchy then there is will more ideas presented and which will lead to the growth of the start-up. And where there are power ideas drifts and there will be only power and fear.

Machine learning is utilized for more than only fraud detection, credit scoring, and other predictions. But they are also employed in the hardware sector, just as they are in the semiconductor sector. Since semiconductors are employed to identify anomalies (Mehta et al.; 2020). Technology CAD is utilized as the training dataset, the enhanced ML model is overfitted, and hidden layers are formed in the simulation data. Additionally, machine learning techniques are utilized in the chemical industry to determine the effectiveness of the cells used in automobiles and how well these automobiles produce batteries for longer battery life (Wang et al.; 2020).

A single individual cannot launch a startup. It must be made by a group. A group of individuals can brainstorm an innovative concept that can be implemented to make it a success. One of the reasons that start-ups thrive is that teams get down and brainstorm ideas, pointing out the benefits and drawbacks. However, not all new businesses succeed; in fact, there are more failures than triumphs. 90% of new firms failed in 2019, according to estimates. In addition, around 21.5% of new businesses fail in their first year on the market, 30% struggle in their subsequent year, 50% struggle in their 5th year, and over 70% fall flat in their 10th year (Bangdiwala et al.; 2022a). As a result, it is difficult for today's entrepreneurs to have the courage to get their start-ups back on track. There is no such thing as a success guidebook or a guaranteed path to success. Everything is dependent on the IPO, stakeholders, concepts, and how the product is presented in the market.

# 3 Research Methodology

In this research project, CRISP-DM is being implemented as depicted in the figure. A CRISP-DM model is used in data mining, business comprehension, and assessment. The CRISP-DM approaches have been used in a lot of data mining applications (Schröer et al.; 2021). The results of this study project served as the basis for this technique. The next sections of the study will go into more depth about this process.

## 3.1 Business Understanding

It will be challenging to manually forecast a start-up's success or for equity investors to make a profitable investment. Additionally, it might be challenging to take into account every aspect that will have an impact on a start-up's success. Additionally, manual decision-making takes a long time and is imprecise. However, when using manual decision-making processes, data mining approaches help take into account all the aspects that are

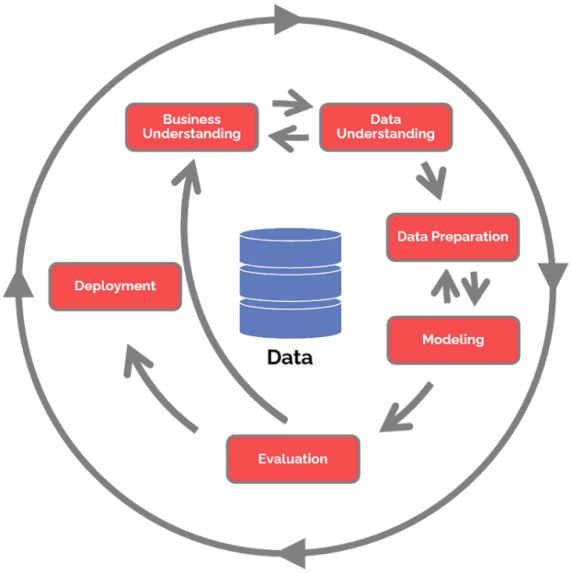


Figure 1: CRISP-DM

concealed from view. Therefore, developing a machine learning model can allow start-up owners, venture capitalists, and angel investors to have less work to do. Determining the importance of Twitter tweets in the success of start-ups is the main goal of this approach. Customers today have a big impact on start-ups as they are the main link between start-ups and venture capitalists or angel investors. Machine learning approaches, as mentioned in the literature review, can aid in the prediction of start-up success and provide clearer views of the factors that influence success.

## 3.2 Data Selection

Two datasets were employed for this research endeavour; one was scraped from Twitter and the other came from Kaggle. The dataset, which comes from the open-source Kaggle platform, is made up of start-ups from the AngelList, which includes details on American start-ups. Even though Twitter offers an API to scrape data, the Search API and API timeline have limited scraping restrictions, and in 6 to 9 days, only a maximum of 3200 tweets may be scraped according to the timeline. Utilizing BeautifulSoup and the Selenium Webdriver to scrape Twitter tweets was therefore simpler. Direct searches of Twitter for the firms included in the AngelList dataset are used to collect the tweets.The typical Series A financing is 6 million USD, Series B is around 11 million USD, Series C is about 16 million USD, and Series D is about 7 million USD. The graph is obviously right-skewed when scaled from 0 to 100 million USD.

## 3.3 Data Pre-Processing

The number of retweets, likes, and the time the tweet was posted is some examples of the information derived from Twitter metadata. The company's and the fundraising rounds' means and standard deviations. Using the dates that were generated, calculate the interquartile range and the dates of the ranges that are scraped. The amount of tweets is high based on the dates, and they are posted every minute.

| | Description | Market | Names | No_Stage_Amount | No_Stage_Date | Pitch | Seed_Amount | Seed_Date | Series_A_Amount | Series_A_Date |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | Cable | Epic-Sciences | | | | | | | |
| 1 | | All Students | Apreso-Classroom | | | | | | | |
| 2 | Visualead (视觉码) creates better interactions be... | Bridging Online and Offline | Visualead | | | Effective and Secure Offline to Mobile experie... | $750,000 | Mar 25, 2012 | $1,600,000 | Aug 15, 2013 |
| 3 | | Food Processing | Onshift | $7,000,000 | Feb 3, 2014 | | | | | |

Figure 2: Funding Dataset

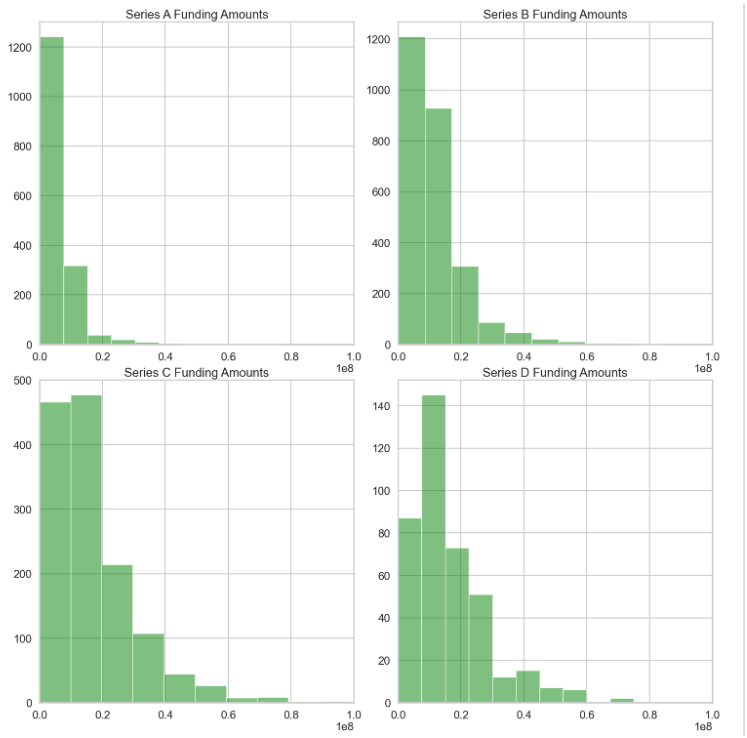


Figure 3: Funding Rounds before applying log function

After the log function is applied to the funding the graphs are less skewed in Figure 4.

## 3.4 Data Transformation

Feature extraction is one of the most crucial processes in the process of data mining since it allows for the progression to the next step, model selection. The process's most crucial and inventive step is this one. This procedure is crucial because it eliminates all the noisy data, which facilitates prediction. The dataset will become smaller as a result. The analysis's forecast will be impacted by the noise data. Additionally, the dataset is split into 80:20 for train and test data.

## 3.5 Modelling

- **Natural Language Processing (NLP):** Using the Python library, the adjectives and nouns from the texts were retrieved. from the Stop Words and Punctuation

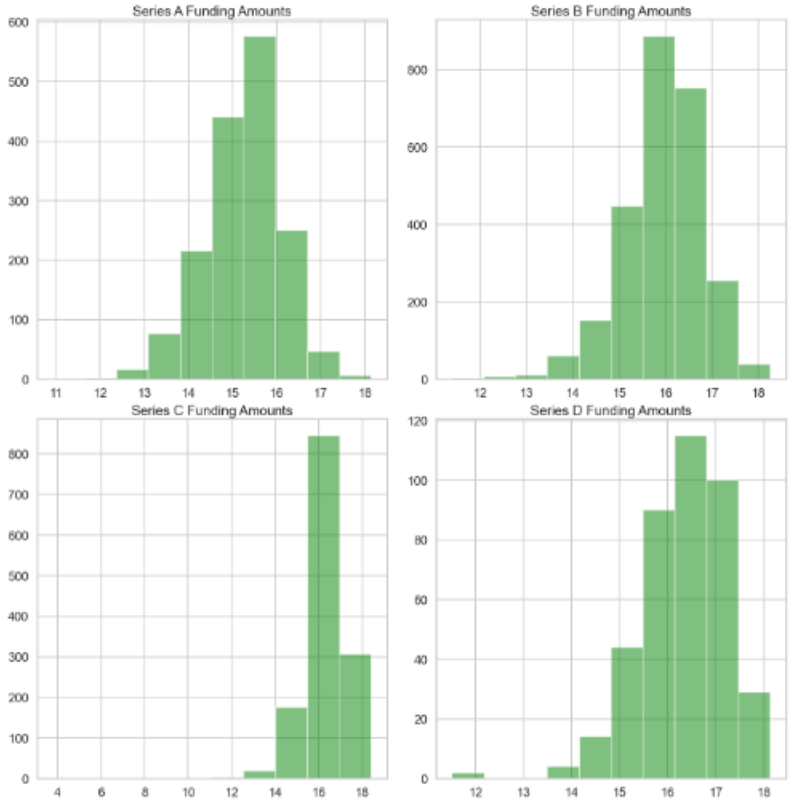


Figure 4: Funding Rounds after applying log function

section of the Sklearn library. The scraped data was used to extract the text. Sentences of text are retrieved in tokenized form from the scraped data. The terms are then simplified to become "talk," "talking," "talked," and so on. It is challenging to separate something since each tweet is limited to 140 characters.

- **Sentiment Analysis:** SentiWordNet 3.0 is utilized here for the sentiment analysis approach since it is an emotion dictionary and has three sentiment scores (0, 1 and 2) that correspond to "positive," "negative," and "neutral" feelings, respectively. Each word in the text is labelled with one of these scores. The parameters employed in the sentiment analysis for each corporate tweet are the mean positive and negative as well as the positive and negative count. Each company's average daily tweet volume is computed.

- **Correlation Analysis:** There were over 150 tweets, and over 1200 firms and the financing data dates corresponded to the study tweet's date. The graph plots the total amount raised (unscaled and scaled) against each numerical characteristic in the table. With the use of this correlation analysis, this can determine how much of an impact each attribute has on the model.

- **Principle component analysis (PCA):** In PCA, the dataset is taken into account as the test, train, and validation datasets without being differentiated. This analysis can assist in determining the linear relation between the dataset's characteristics and assist in making precise predictions. Additionally, it will aid in understanding the variances in the data being gathered. Insightful PCA will assist

in reducing the predominance of one component over another so that it won't affect the variance as a whole. To standardize the variation of the whole analysis, the Box-Cox transformation is applied to all characteristics that are present in all columns. To normalize the data, the funding log is used there. Tweets, likes, and retweets are among the most important components that make up the variance, and according to PCA analysis, there are just 5 characteristics that demonstrate the majority of the variance, and in that, 3 characteristics cover around 98% of the variation.

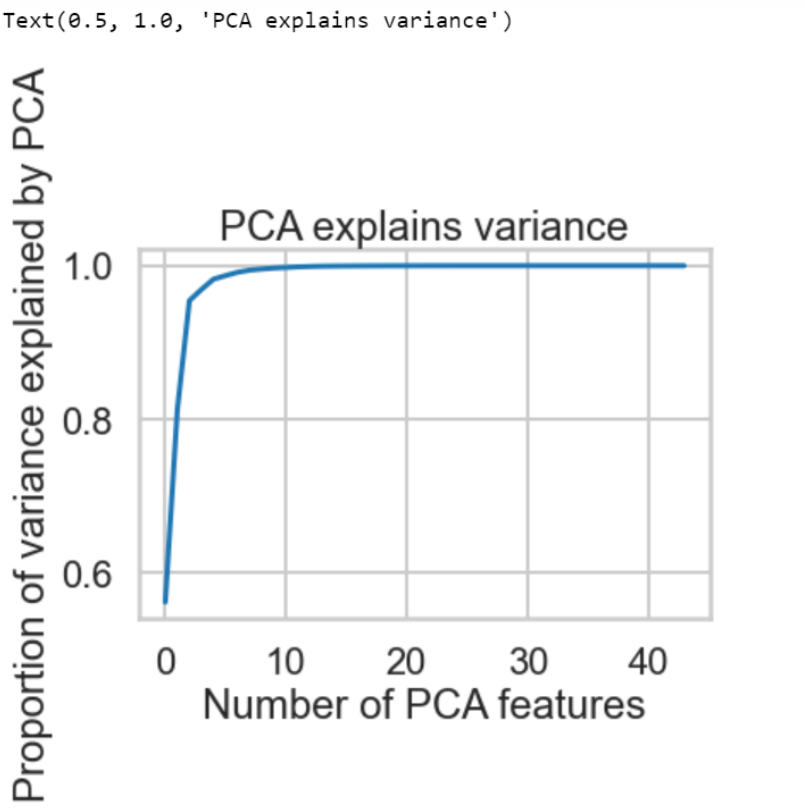


Figure 5: Principle Component Analysis (PCA)

- **K-fold Cross-Validation:** By splitting the dataset into the validation/test and train datasets, the K-fold statistical analysis takes into account all the numerical features of the two datasets. By doing this, the model is trained five times, and each fold's validation is carried out separately to acquire the hyperparameters. Additionally, unmodified test data was used to assess dependability. The K-fold standard statistical assists in improving model predictions.

- **Baseline Prediction (Naive Bayes):** The Naive Bayes approach was utilized for the baseline model evaluation, and this prediction will be used to evaluate the other models that will be used. This will enable the creation of comprehensive and accurate predictions.

- **Linear Regression:** Two logistic regressions called Lasso and Ridge were utilised to conduct the standard Linear regression on the scaled data. These two methods

restrict the parameter vectors using the L-1 and L-2 norms.

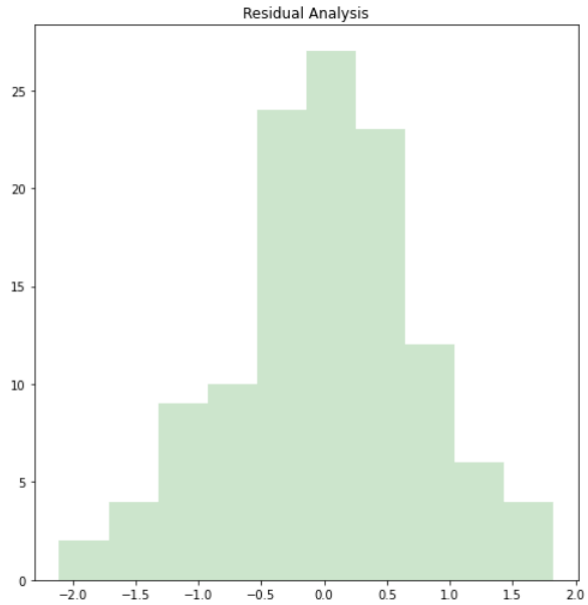


Figure 6: Lasso Regression Residual Analysis

Figure 6 shows the residual analysis of Lasso Regression. And Figure 7 shows the residual analysis of the Ridge regression. Both figure show normalization and the features used for this analysis were unscaled.

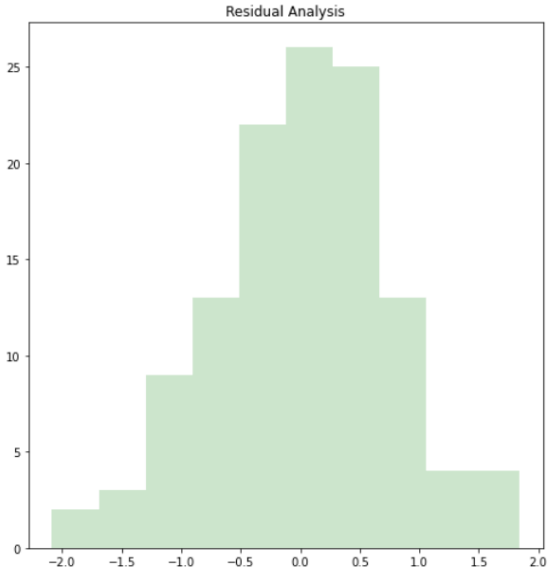


Figure 7: Ridge Regression Residual Analysis

- **Sentiment Analysis:** The SentiWordNet 3.0 dictionary was utilized for the adjectives and nouns in this study endeavour, and three sentiment ratings were employed: neutral, positive, and negative. The mean positive and mean negative for each tweet was computed. The positive and negative sentiments are defined as if the positive is more than 0.5 and the negative is greater than 0.5. Each tweet is labelled as 0, 1, or 2. To summarize, four characteristics are employed in sentiment analysis: positive count, average positive, negative count, and average negative. To determine the correlation with the financing round, the average of the four attributes for the tweets evaluated is used.

- **Support Vector Regression (SVR):** The three employed were the polynomial, linear, and RBF kernel choices. using the use of 5 K-fold analysis cross-validation, and the GridSearchCV. As a result, the optimal parameters for the analysis have been found. The RMSE of each simulation will then be evaluated and compared, and the model with the lowest RMSE will be considered the best model. Figure 8 shows the PCA analysis scatter plot and the features are spread throughout the graph.

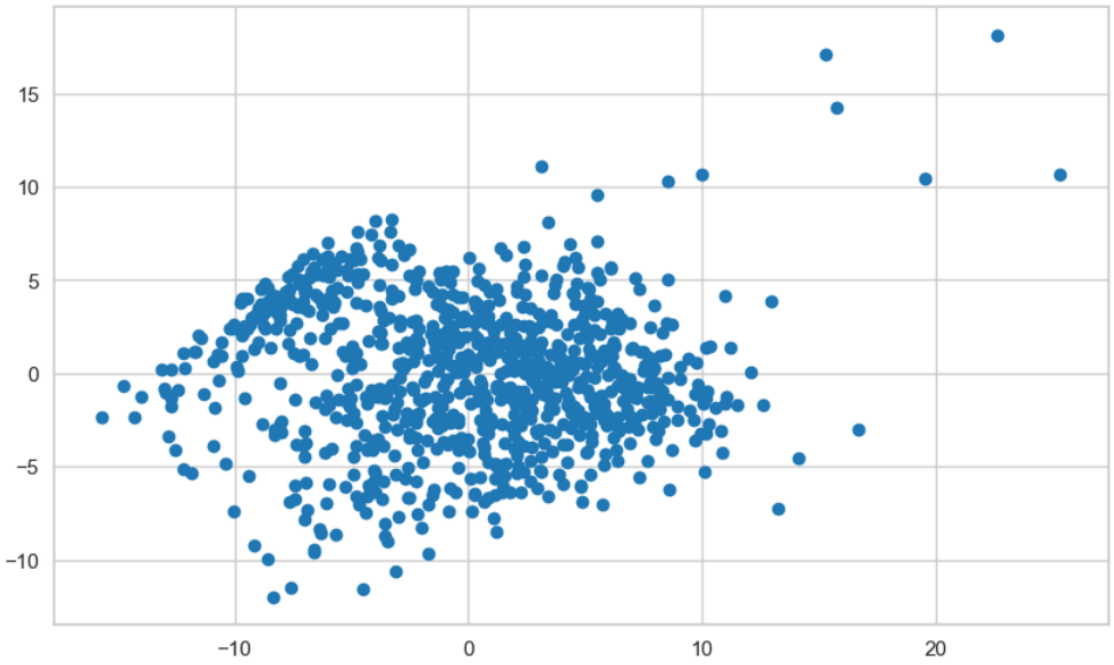


Figure 8: PCA analysis scatter plot

- **Random Forest Regression:** In the random forest, regression is similar to an algorithm that integrates all of the model predictions and produces superior results when they are all merged.

- **Recurrsion Neural Network (RNN):** The results of the nodes in a feed-forward network like an RNN might influence the input to the next nodes. With text analysis, sentiment analysis, handwriting recognition, and other tasks, the RNN performs well.

# 4 Design Specifications

This section includes a discussion of the techniques' broad use as well as an analysis of their results. This section covers the setup process, data modelling, and model building. It also mentions the tools that were used to perform the assignment. For the study endeavour, a Microsoft Windows 10 system with 64-bit, 16GB of RAM, and a 500GB hard drive were employed. And to store the data, a 500GD SSD was employed.

## 4.1 Environment Setup

- **Python packages:** Python packages are easy to use and have a lot of capability for data mining techniques. It includes data analytics and machine learning technologies that are essential for execution. The main benefit of using Python is this. The packages are used for many tasks, including data modelling, machine learning, and data cleansing.
- **Jupyter Notebook:** The most recent version of Python is compatible with this IDE, which consumed the fewest resources feasible.

## 4.2 Data Transformation

Two datasets were employed in this research project's data extraction method so that it could be analyzed in accordance with the model's specifications. The Selenium Webdriver and BeautifulSoup were used to scrape the Twitter dataset from the social media site. 13 columns make up the Kaggle dataset, which is utilized for model analysis. Since the data that is extracted from Twitter is not in a format that allows for model-based analysis, from the scripts the data was extracted. the characteristics required for the analysis. The data is taken from Twitter features and saved in a CSV file for further usage. The study was conducted on around 300,00 tweets because of the restricted source availability, and the models take a long time to run when the data is too large for analysis. Duplicate and null values were removed from both the test dataset and the training dataset. In this study, the model construction process was broken down into three distinct steps, the first of which included the LDA model (Saura et al.; 2019). Sentiment analysis and text analysis were conducted on the Twitter comments with the hashtags that were retrieved during this modelling step.

## 4.3 Naive Bayes

The Naive Bayes technique was used to assess the baseline model, and this prediction will be used to analyze the other models that will be deployed. This will allow for the development of thorough and reliable predictions. In this research, the Naive Bayes model serves as the baseline prediction for all other models. The RMSE of the models is predicted and then compared to the other models. The train and the test dataset were split into 80:20 respectively for the model evaluation. The RMSE obtained for Naive Bayes is 99.69. As this is a baseline prediction the RMSE value will be higher due to the outliers present in the dataset.

## 4.4 Support Vector Regressor (SVR)

The polynomial, linear, and RBF kernels were employed. GridSearchCV and cross-validation with 5 K-fold analysis were employed. As a result, the optimum study parameters have been found. The RMSE of each simulation will then be examined and compared, and the model with the lowest RMSE will be proclaimed the best model. SVR RBF, SVR polynomial, and SVR linear have RMSE values of 0.995, 0.998, and 0.997, respectively.

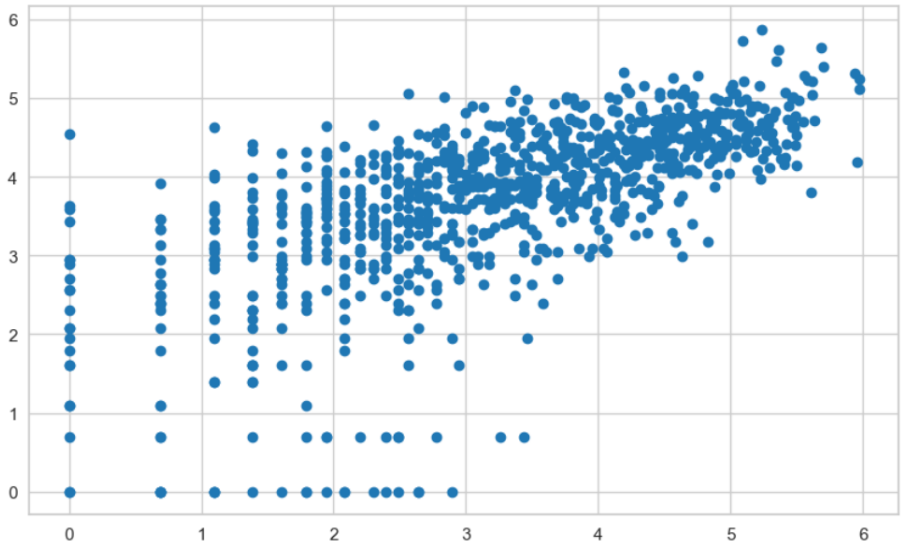


Figure 9: Graph of likes vs retweets

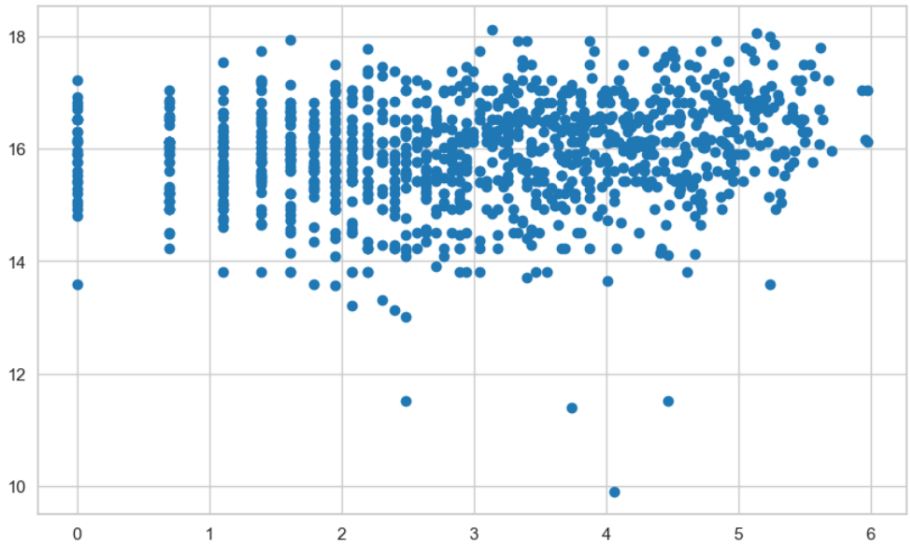


Figure 10: Graph of likes vs series amount

Figures 9 and 10 are the graphs of the plotted likes vs. retweets and likes versus the series amount.

## 4.5 Linear Regression (LR)

Two logistic regressions, Lasso and Ridge, were employed to do the standard Linear regression on the scaled data. These two techniques confine the parameter vectors by using the L-1 and L-2 norms. The nature of Lasso and Ridge Regression is extremely similar. The magnitude of the features is considered in Lasso regression, whereas the square of the magnitude is considered in Ridge regression. The RMSE value for the Lasso regression and Ridge regression is 7.4 and 7.5 respectively.

## 4.6 Sentiment Analysis

The SentiWordNet 3.0 dictionary was employed in this study endeavour for the adjectives and nouns, and three sentiment ratings were used: neutral, positive, and negative. The mean positive and mean negative for each tweet was determined. The positive and negative sentiments are defined here as if the positive is more than 0.5 and the negative is greater than 0.5. Each tweet has a label of 0, 1, or 2. To sum up, the sentiment analysis employs four features: positive count, average positive, negative count, and average negative. The average of the four attributes for the tweets evaluated is used to discover the correlation with the fundraising round. And the average positives was 89%. So from this, we can say that most of the comments were positive.

## 4.7 Random Forest Regression

The Random Forest method integrates all of the Decision Tree findings to predict or categorize the variable. To eliminate discrepancies in the correlation of several decision trees. As the trees are given training data, they grow. Various subnets are generated along the process, which is known as bagging (Rodriguez-Galiano et al.; 2015). The RF assists the tree in growing by randomly selecting the best feature from the subnet and feeding that feature to the next layer of the tree, resulting in the best fit. The dataset that was not utilized in training is used at the conclusion of the analysis. And the RF tweaks one feature at a time to observe if the model's accuracy increases or decreases. The RMSE of the Random forest is 1.9.

## 4.8 Recursive Neural Network (RNN)

The dataset was divided into 80:20 where 175,216 tweets were used for training and 84,420 tweets were used for testing the model during the RNN implementation. The Keras Python library was utilized in this model because it functions as an interface between Python and TensorFlow. This model attained an accuracy of 81%. And 50 epochs were completed, which is the number of times the step was iterated. And the RMSE value for the RNN is 0.65.

# 5 Evaluation

Many models were utilized in this study effort for more precision, and the RMSE values were used to compare the 7 models. And each statistic has benefits and drawbacks. The Root Mean Squared Error was calculated using the equation below (RMSE). The primary purpose is to forecast the influence of tweets on funding data. The correlation research revealed that there is a correlation between the tweets and the funding feature. Because funding is the dependent variable, a log to funding was applied to obtain a normalization function. After applying the log function, there were five characteristics with the highest correlation, with three of them showing an overall 98% correlation. Following the correlation study, the Nave Bayes baseline prediction was used, yielding an RMSE of 99.69. Then, for parameter vectorization, linear regressions Lasso and Ridge regression were employed on scaled and unscaled data in the L1 and L2 forms. Lasso and Ridge have RMSE values of 7.4 and 7.5, respectively. Then, for parameter vectorization, linear regressions Lasso and Ridge regression were employed on scaled and unscaled data in the L1 and L2 forms. Lasso and Ridge have RMSE values of 7.4 and 7.5, respectively. There were three types of kernels used in support vector regression: linear, polynomial, and RBF. Each algorithm's RMSE value was calculated and compared. The RMSE values obtained were 0.998, 0.995, and 0.997, respectively. The random forest technique was used, and the RMSE was 1.9. The RMSE for the RNN algorithm is 0.65. In conclusion, RNN outperformed the other algorithms.

# 6 Conclusion and Discussion

Many challenges were faced at the start of the research, and many of these were dealt with and resolved. There are hidden dependencies in the tweets that greatly affect the correlation. Here the importance was given to the tweets, likes, retweets and others. The K-fold statistical method was used for the hyperparameter tuning. In this research, all the other algorithms performed better than the baseline prediction but RNN outperformed. And tweets are one of the contributing factors to the success of funding start-ups. The tweets, however, include underlying meanings and connotations that were not covered here.

Table 1: A table caption.

| **Model** | **RMSE** |
|---|---|
| Naive Bayes | 99.69 |
| Lasso Regression | 7.4 |
| Ridge regression | 7.5 |
| SVR RBF | 0.995 |
| SVR Polynomial | 0.998 |
| SVR linear | 0.997 |
| Random Forest Regressor | 1.9 |
| Recursive Neural Network (RNN) | 0.65 |

# 7 Future Work

During the investigation, seven ML models were utilized and compared. More than tweets, there are several other elements that influence the success of start-ups. Some significant enhancements that may be added to this research include:

- **Clustering:** More clustering could be undertaken to provide personalized projections by taking into account the size of the organization and integrating other parameters with Twitter data to achieve better results.
- **Hyperparameter Tuning using Keras:** We can optimize hyperparameters using Keras tuning. We can gain better visualization tools for better predictions as a result of this. And the inaccuracy and outliers are readily dealt with here.

# References


Ang, Y. Q., Chia, A. and Saghafian, S. (2022). Using machine learning to demystify startups' funding, post-money valuation, and success, *Innovative Technology at the Interface of Finance and Operations*, Springer, pp. 271–296.

Arroyo, J., Corea, F., Jimenez-Diaz, G. and Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments, *Ieee Access* **7**: 124233–124243.

Åstebro, T. and Bernhardt, I. (2003). Start-up financing, owner characteristics, and survival, *Journal of Economics and Business* **55**(4): 303–319.

Bangdiwala, M., Mehta, Y., Agrawal, S. and Ghane, S. (2022a). Predicting success rate of startups using machine learning algorithms, *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, pp. 1–6.

Bangdiwala, M., Mehta, Y., Agrawal, S. and Ghane, S. (2022b). Predicting success rate of startups using machine learning algorithms, *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–6.

Buchanan, B. G. and Wright, D. (2021). The impact of machine learning on uk financial services, *Oxford Review of Economic Policy* **37**(3): 537–563.

Corea, F., Bertinetti, G. and Cervellati, E. M. (2021). Hacking the venture industry: An early-stage startups investment framework for data-driven investors, *Machine Learning with Applications* **5**: 100062.

de Souza Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I. and Conte, T. (2019). Understanding development process of machine learning systems: Challenges and solutions, *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, IEEE, pp. 1–6.

Gidron, B., Israel-Cohen, Y., Bar, K., Silberstein, D., Lustig, M. and Kandel, D. (2021). Impact tech startups: A conceptual framework, machine-learning-based methodology and future research directions, *Sustainability* **13**(18): 10048.

Hopkins, A. and Booth, S. (2021). Machine learning practices outside big tech: How resource constraints challenge responsible development, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 134–145.

Lee, S. (2022). The myth of the flat start-up: Reconsidering the organizational structure of start-ups, *Strategic Management Journal* **43**(1): 58–92.

Mayer, S. (2022). Financing breakthroughs under failure risk, *Journal of Financial Economics* **144**(3): 807–848.

Mehta, K., Raju, S. S., Xiao, M., Wang, B., Zhang, Y. and Wong, H. Y. (2020). Improvement of tcad augmented machine learning using autoencoder for semiconductor variation identification and inverse design, *IEEE Access* **8**: 143519–143529.

Pasayat, A. K. and Bhowmick, B. (2021). An evolutionary algorithm-based framework for determining crucial features contributing to the success of a start-up, *2021 IEEE Technology Engineering Management Conference - Europe (TEMSCON-EUR)*, pp. 1–6.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, *Ore Geology Reviews* **71**: 804–818.

Saura, J. R., Palos-Sanchez, P. and Grilo, A. (2019). Detecting indicators for startup business success: Sentiment analysis using text data mining, *Sustainability* **11**(3): 917.

Schröer, C., Kruse, F. and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model, *Procedia Computer Science* **181**: 526–534.

Sharma, U. and Manchanda, N. (2020). Predicting and improving entrepreneurial competency in university students using machine learning algorithms, *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, pp. 305–309.

Tang, Z., Yang, Y., Li, W., Lian, D. and Duan, L. (2022). Deep cross-attention network for crowdfunding success prediction, *IEEE Transactions on Multimedia* pp. 1–1.

Veloso, F. (2020). *Predicting Startup Success in US*, PhD thesis, The University of North Carolina at Charlotte.

Vijai, C. and Wisetsri, W. (2021). Rise of artificial intelligence in healthcare startups in india, *Advances In Management* **14**(1): 48–52.

Wang, Y., Seo, B., Wang, B., Zamel, N., Jiao, K. and Adroher, X. C. (2020). Fundamentals, materials, and machine learning of polymer electrolyte membrane fuel cell technology, *Energy and AI* **1**: 100014.

Weber, M., Beutter, M., Weking, J., Böhm, M. and Krcmar, H. (2022). Ai startup business models, *Business & Information Systems Engineering* **64**(1): 91–109.

Yin, D., Li, J. and Wu, G. (2021). Solving the data sparsity problem in predicting the success of the startups with machine learning methods, *arXiv preprint arXiv:2112.07985* .

Żbikowski, K. and Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using crunchbase data, *Information Processing & Management* **58**(4): 102555.